Assessing Demographic Fairness in Deep Learning for Glaucoma Diagnosis Using 2D OCT RNFL Images

Shrey Jaiswal[1] and Gaurav Ghosh[2] and Vanshita Mehta[3]

[1]BML Munjal University, Kapriwas, Haryana
shrey.jaiswal.23cse@bmu.edu.in
[2]BML Munjal University, Kapriwas, Haryana
gaurav.ghosh.23cse@bmu.edu.in
[3]BML Munjal University, Kapriwas, Haryana
vanshita.mehta.23cse@bmu.edu.in

Abstract. Glaucoma is a leading cause of irreversible blindness, and early detection using optical coherence tomography (OCT) retinal nerve fiber layer (RNFL) imaging has become a critical component of modern clinical workflows. However, recent studies have raised concerns regarding demographic bias in deep learning models trained on ophthalmic datasets. In this study, we develop and evaluate a deep learning–based glaucoma classification framework using 2D OCT RNFL images from the Harvard Glaucoma Fairness dataset. The proposed model achieves a test accuracy of 0.7678, precision of 0.8153, recall of 0.7403, and an F1 score of 0.776. Fairness evaluation demonstrates strong overall discriminatory performance (AUC = 0.8538 and ES-AUC = 0.75) with variation across racial groups: 0.8539 (Asian), 0.8106 (Black), and 0.8511 (White). We further quantify fairness using Demographic Parity Difference (DPD = 0.1333) and Difference in Equalized Odds (DE Odds = 0.831), revealing measurable but moderate bias. These findings highlight the need for fairness-aware optimization when deploying automated glaucoma screening models in diverse populations. Our work establishes a baseline analysis for future research on equitable AI-driven ophthalmic diagnostics.

Keywords: Glaucoma detection, Optical coherence tomography (OCT), Retinal nerve fiber layer (RNFL) imaging, Deep learning, Algorithmic fairness, Demographic bias, Demographic parity.

1 Introduction

Glaucoma is one of the leading causes of irreversible blindness worldwide, affecting more than 76 million individuals, with prevalence projected to increase substantially in the coming decades [1]. Early diagnosis is essential for preventing vision loss, as structural changes to the optic nerve head and retinal nerve fiber layer (RNFL) often occur before noticeable functional impairment. [2] Optical coherence tomography (OCT) has emerged as a widely adopted imaging modality for quantifying RNFL thickness and detecting glaucomatous damage. In parallel, recent advances in deep learning have accelerated the development of automated systems capable of interpreting OCT scans with accuracy comparable to expert clinicians. [3]

Despite these advancements, concerns regarding algorithmic fairness have become increasingly prominent in medical AI research. Several studies have shown that deep learning models trained on imbalanced or non-representative datasets can exhibit performance disparities across demographic subgroups. [4] Such disparities are particularly problematic in ophthalmology, where disease prevalence, ocular structure, and imaging characteristics may vary between racial and ethnic populations. As a result, a glaucoma detection model that performs well on the overall population may still underperform for specific subgroups, potentially exacerbating existing health inequities.

In this study, we develop a deep learning model for glaucoma detection using 2D OCT RNFL images from the Harvard Glaucoma Fairness Dataset, a large and demographically diverse dataset specifically designed for bias analysis in ophthalmic AI. [5] Our work has two objectives:

1. To evaluate the diagnostic performance of a convolutional neural network for RNFL-based glaucoma classification, and;
2. To quantify demographic fairness across three racial (Asian, Black, and White) population groups using standard fairness metrics.

The proposed model demonstrates strong overall predictive performance while revealing measurable disparities between demographic subgroups. Through a structured analysis of both accuracy and fairness measures, this study provides a foundational benchmark for future research aimed at developing equitable, trustworthy, and clinically deployable glaucoma-screening systems across racial groups. Ultimately, our findings emphasize the importance of integrating fairness evaluation into the development pipeline of medical AI models, especially those intended for diverse real-world patient populations. [6]

2 Related Work

Deep learning–based glaucoma detection has been extensively investigated across multiple ophthalmic modalities, including OCT volumes, RNFL thickness maps, and retinal fundus photographs. Despite rapid methodological advances, most prior work offers limited consideration of fairness across demographic groups. Early OCT-based glaucoma models, such as those by Erfan Noury et al. (2022) [7] and Maetschke et al. (2019) [8], demonstrate strong diagnostic accuracy but are trained on datasets containing little to no demographic metadata. Consequently, these models are unable to assess or mitigate demographic disparities, and their evaluation protocols do not report race, ethnicity, or gender specific performance. Even studies focusing on interpretability through saliency maps or biologically guided feature visualization fail to address potential demographic biases arising from structural variations or differences in disease prevalence.

More recent multimodal and semi-supervised approaches, including those based on the Harvard Glaucoma Detection and Progression dataset by Luo et al. (2023) [9], expand model capacity yet continue to provide limited fairness analysis. These methods often rely on pseudo-labeling or policy networks to improve generalization but lack explicit mechanisms for detecting or correcting cross-group imbalances. Fundus-based models proposed by Jisy et al. (2024) [10], Pascal et al. (2022) [11], and Bajwa et al. (2019) [12] exhibit similar limitations: they depend heavily on supervised annotations, frequently draw from geographically or ethnically narrow datasets, and rarely report subgroup-wise metrics. As a result, many existing models remain susceptible to demographic biases embedded within the training data.

A significant advancement in this domain is the Harvard Glaucoma Fairness (Harvard-GF) dataset [5], introduced by Luo et al. (2024). Harvard-GF represents the first medical imaging dataset explicitly designed for fairness research with 2D and 3D image data. It offers balanced representation across major racial groups (Asian, Black, White), extensive demographic attributes, and both 2D RNFLT maps and 3D OCT B-scans. In conjunction with the dataset, the authors propose Fair Identity Normalization (FIN), a demographic-aware feature-level normalization method that learns group-specific means and variances to equalize feature importance across racial and gender groups. Their results demonstrate substantial improvements in equity-scaled performance metrics, with FIN outperforming several state-of-the-art adversarial and contrastive fairness baselines.

Although Harvard-GF addresses the critical need for equitable datasets in medical imaging, most existing glaucoma-detection methods—including FIN—depend on fairness-specific modules or task-specific debiasing losses to mitigate disparities. Furthermore, FIN introduces explicit demographic conditioning into the feature extraction process, which may be unsuitable in contexts where demographic information cannot be used, is incomplete, or where models are required to remain demography-agnostic.

In contrast, the present work investigates whether fairness and subgroup robustness can be improved without identity-conditioned normalization or fairness-specific loss functions. We explore whether architectural regularization strategies—such as dropout, stochastic depth, and Gaussian noise—can implicitly enhance demographic equity while simultaneously improving overall diagnostic accuracy. Using the Harvard-GF dataset for evaluation and comparing directly against FIN and FSCL+ w/ FIN, our study examines a complementary direction: achieving fairness not through explicit demographic conditioning, but through stronger generalization and more stable learned representations.

3 Methodology

Within this section, the authors have discussed in depth the methods applied to achieve the object of this study. The model architecture is a modified version of the ResNet18 model which leverages multiple regularization techniques to ensure generalizability to unseen data and ensure that the model learns high level representative features. In section 3.2, the evaluation metrics devised to judge the classification and fairness performance have been discussed.

3.1 Model Architecture

We adopt a modified ResNet-18 [13] architecture tailored for 2D peripapillary OCT RNFL imaging and enhanced with multiple regularization mechanisms to improve generalization and reduce overfitting. The model, termed **StochDepth-ResNet18**, is based on the original ResNet-18 convolutional backbone but incorporates a 1-channel input interface, stochastic depth applied at the level of individual residual blocks, and dropout after each residual stage and before the final classification layer. Each modification is explained below.

Standard ResNet-18 is designed for 3-channel RGB images. Because OCT RNFL scans are single-channel grayscale images, the first convolutional layer is replaced with:

Conv2d (1→64, 7×7, Stride=2, Padding=3)

This change allows the network to learn feature representations directly from structural retinal thickness patterns without redundant colour filters. The remaining backbone components, BatchNorm, ReLU, max-pool, and the four residual stages (layer1 - layer4), are retained from the original ResNet-18. Each stage contains two BasicBlock residual units, resulting in a total of 8 residual blocks across the network.

A ResNet BasicBlock consists of two sequential 3 × 3 convolutions with Batch Normalization and ReLU activation. A skip connection adds the input feature map $\boldsymbol{x}$ to the output of the two convolutions:

$$output \ = \ ReLU \ (F(x) \ + identity(x)) \tag{1}$$

where *F(x)* denotes the two-layer convolutional path and the identity projection is optionally downsampled whenever the spatial resolution or number of channels changes. Residual connections improve gradient flow, enabling deeper and more stable models.

A key modification is the integration of layer-wise stochastic depth [14], where each residual block is randomly dropped during training. For block index $i \in \{0, 1, \ldots, 7\}$ out of 8 total blocks, the drop probability is linearly scaled:

$$p_{stoch} \ = \frac{i}{N} \cdot p_{max} \ where \ p_{max} \ = \ 0.1 \tag{2}$$

During training, with probability p_i, the residual branch output is either preserved or dropped according to:

$$\tilde{F}(x) \ = \begin{cases} \quad 0, & with \ probability \ p_i \\ F(x), & else \end{cases} \tag{3}$$

To preserve expected feature magnitude, surviving paths are rescaled by $1/(1 - p_i)$. At inference time, stochastic depth is disabled, and all blocks remain active. This mechanism forces the model to remain robust to the removal of deep layers and reduces reliance on any single residual block, thereby improving generalization on small medical datasets and reducing sensitivity to demographic-specific patterns.

Following each of the four ResNet stages, the feature maps are regularized using spatial dropout with probability $p = 0.5$:

$$x \leftarrow Dropout(x),\ after\ layer1,\ layer2,\ layer3,\ layer4.$$

This dropout is applied once per stage, not per block. It reduces co-adaptation of stage-level features and provides an additional regularization pathway complementary to stochastic depth.

After processing all residual stages, the network applies global average pooling and flattening to compress the spatial feature maps:

$$GAP: (512 \times H \times W) \rightarrow (512).$$

The resulting 512-dimensional feature vector is passed directly to a linear layer:

$$FC(512 \rightarrow 1),$$

which outputs a single logit representing the predicted probability of glaucoma. No dropout is applied after global pooling in the final version of the model. Fig. 1 depicts the architecture of the StochDepth-ResNet18 model with the implementations of the dropout and stochastic depth regularizations.

3.2 Evaluation Metrics

To quantitatively assess model performance and demographic fairness, we employed a combination of standard binary-classification metrics and group-aware fairness measures. All metrics were computed on the held-out test set, using the model outputs prior to thresholding as probability estimates.

3.2.1 Classification Metrics. Because the model outputs a single logit, $\hat{\mathbf{z}}$, the corresponding probability is obtained via the sigmoid function:

$$\hat{p} = \sigma(\hat{z}) = \frac{1}{1 + e^{-\hat{z}}} \tag{4}$$

Binary predictions are produced by applying a fixed threshold as discussed in (5).

$$\hat{y} = \begin{cases} 1, & \hat{p} \geq 0.5, \\ 0, & \hat{p} < 0.5. \end{cases} \tag{5}$$

The following performance metrics were then computed:

Accuracy.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall (Sensitivity).

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

Specificity.

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

F1-Score.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

ROC-AUC. The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was computed using the unthresholded probabilities $\hat{p}$, following the standard trapezoidal integration implemented in scikit-learn.

Confusion Matrix. Given predicted labels $\hat{y}$ and true labels y, a 2×2 matrix of TP, FP, FN, TN was generated using scikit-learn's confusion_matrix, and displayed using ConfusionMatrixDisplay.

3.2.2 Fairness Metrics. To evaluate demographic fairness, we compute subgroup-wise performance measures and derive composite fairness metrics. Let A denote the sensitive attribute (here, race), and let g index the demographic groups present in the test set. For each demographic group g, we extract the subset D_g of samples satisfying A = g. Groups with fewer than 5 samples are skipped for stability and for each included subgroup, we compute:

AUC.

$$AUC_g = ROC\_AUC(y_g, \hat{p}_g) \tag{11}$$

Accuracy.

$$Accuracy_g = \frac{TP_g + TN_g}{TP_g + TN_g + FP_g + FN_g} \tag{12}$$

True Positive Rate.

$$TPR_g = \frac{TP_g}{TP_g + FN_g + 10^{-8}} \tag{13}$$

where the small constant $10^{-8}$is added exactly as in the implementation to avoid division by zero.

False Positive Rate.

$$FPR_g = \frac{FP_g}{FP_g + TN_g + 10^{-8}} \tag{14}$$

Positive Prediction Rate. This corresponds to the mean predicted label in the group:

$$\pi_g = E[\hat{y} \mid A = g] \tag{15}$$

For the complete test set, we compute:

Overall AUC.

$$AUC_{overall} = ROC\_AUC(y, \hat{p}) \tag{16}$$

Overall Accuracy.

$$Acc_{overall} = Acc(y,\ \hat{y}) \tag{17}$$

Let N denote the number of groups that passed the sample-size threshold. The Harvard GF benchmark penalizes performance disparity across demographic groups.

AUC Disparity.

$$\Delta AUC = \frac{1}{N}\Sigma_g | \boldsymbol{AUC_{overall}} - \boldsymbol{AUC_g} | \tag{18}$$

Equity-Scaled AUC.

$$ES - AUC = \frac{\boldsymbol{AUC_{overall}}}{1 + \Delta AUC} \tag{19}$$

Accuracy Disparity.

$$\Delta Acc = \frac{1}{N}\Sigma_g | \boldsymbol{Acc_{overall}} - \boldsymbol{Acc_g} | \tag{20}$$

Equity-Scaled Accuracy.

$$ES - Acc = \frac{\boldsymbol{Acc_{overall}}}{1 + \Delta Acc} \tag{21}$$

Demographic Parity Difference (DPD). DPD measures disparity in the positive prediction rate across groups.

$$DPD = max(\pi_g) - min(\pi_g) \tag{22}$$

Differential Equalized Odds (DEOdds). DEOdds quantifies the average discrepancy in both true positive rates and false positive rates across groups.

$$\text{DEOdds} = \frac{1}{2}[\max(TPR_g) - \min(TPR_g) + \max(\text{F}PR_g) - \min(\text{F}PR_g)] \tag{23}$$

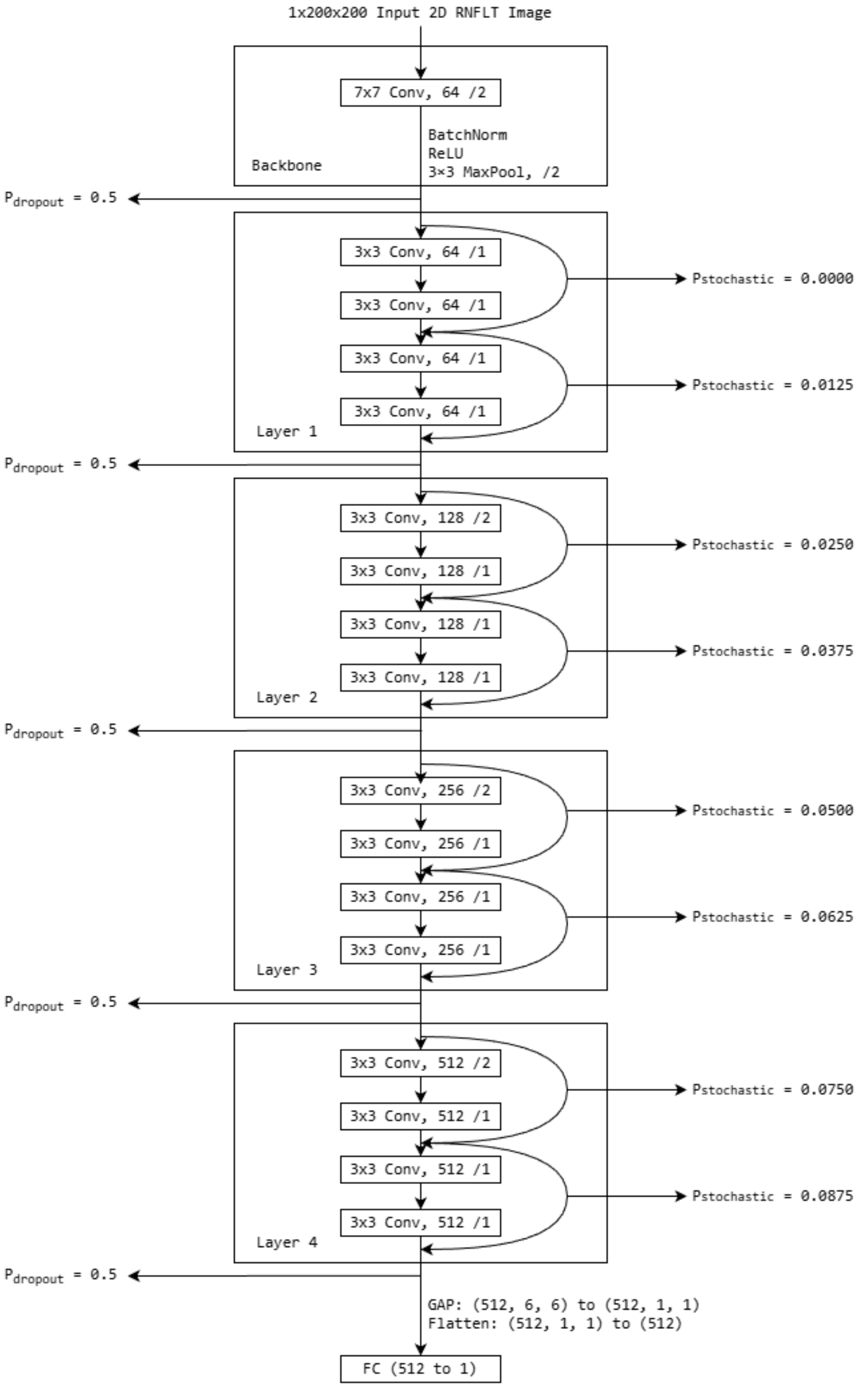

Fig. 1. StochDepth-ResNet18 Model Architecture

4 Experimental Design

This study utilizes the Harvard Glaucoma Fairness (HGF) Dataset [5], a publicly available ophthalmic imaging dataset designed to support research on bias, fairness, and performance disparities in automated glaucoma detection systems. The dataset contains 3,300 patients, each providing a single 2D peripapillary OCT Retinal Nerve Fiber Layer (RNFL) scan, collected under controlled clinical protocols.

The object of this research paper is to develop a glaucoma detection model with a robust and fair performance across racial groups. A crucial precursor to achieving that objective is to have a dataset with a balanced distribution of races. Table 1 shows the distribution of classes in the HGF dataset.

Table 1. Distribution of Races in HGF dataset

Racial Group	Number of Images	Percent of Dataset
Black	1100	33.33%
White	1100	33.33%
Asian	1100	33.33%

As we can observe, an equal distribution across racial groups makes the Harvard GDF dataset suitable for the purpose of this study.

4.1 Dataset Splitting

The dataset provides a pre-assigned "use" label that designates each sample as belonging to the training, validation, or test subset. We adopt this predefined split to ensure consistency with the dataset's intended usage protocol and to prevent any inadvertent overlap of patient-level information across subsets.

Although the splits are predetermined, it is essential to verify that demographic distribution remains balanced across subsets to avoid introducing unintentional biases during model training or evaluation.

Table 2. Distribution of racial groups across splits

Racial Group	Number of Images per Split		
	Training	Testing	Validation
Black	700	300	100
White	700	300	100
Asian	700	300	100

As Table 2 shows, the distribution of classes remains perfectly balanced across each split. Thus, we go forward with the current split technique. The split ratio across training, testing and validation is 70:30:10. A total of 2100 images are used to train the model and its performance is evaluated on a test set of 900 images.

4.2 Training Procedure

The model was trained end-to-end using supervised learning on the training split defined by the dataset's pre-assigned use labels. All experiments were conducted in PyTorch, following a standardized optimization strategy designed to promote stability, mitigate overfitting, and ensure fair generalization across demographic subgroups. Since the task is binary glaucoma classification, the model outputs a single logit, and optimization is performed using the Binary Cross-Entropy loss with logits (BCEWithLogitsLoss), which internally applies the sigmoid function in a numerically stable form. We adopt the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-3} to regularize the parameter magnitudes and reduce overfitting. A StepLR scheduler decays the learning rate by a factor of 0.5 every 10 epochs to ensure stable convergence.

To improve robustness in low-data medical imaging settings, we inject Gaussian noise into the input images during training:

$$x_{noisy} = x + N(0, 0.05^2) \tag{24}$$

This augmentation is applied only in training mode and forces the model to learn stable features invariant to mild perturbations of the RNFL representation. After computing the loss, gradients are backpropagated through the entire network. To prevent instabilities, especially when stochastic depth removes deep residual branches, we apply gradient clipping with a maximum norm of 1.0:

$$||\nabla\theta||_2 \leq 1.0 \tag{25}$$

where $\nabla\theta$ represents the gradients of the model parameters or weights and $||\nabla\theta||_2$ represents the L2 norm (Euclidean length) of the gradient vector. This protects the optimization process from exploding gradients and supports stable convergence.

At the end of each epoch, the model is evaluated on the validation split using the same BCE-with-logits loss and sigmoid-derived metrics (accuracy, precision, recall, F1). No dropout, stochastic depth, or noise augmentation is applied during validation (model.eval()). The learning rate scheduler is stepped once per epoch, adjusting the learning rate according to the predefined schedule. Fig. 3 displays the training versus validation loss across epochs.

5 Results

5.1 Comparative Performance of ResNet 18 Variants

Table 3. Comparative Performance of ResNet 18 Variants

Model	Accuracy	Precision	Recall	Specificity	F1 Score	ROC-AUC
Model 1	0.7578	0.7505	**0.8303**	0.6715	**0.7883**	0.8524
Model 2	0.7633	0.7749	0.7955	0.7251	0.7851	0.8440
Model 3	**0.7678**	**0.8153**	0.7403	**0.8005**	0.7760	**0.8538**

Table 3 presents the comparative evaluation of the three models across key diagnostic metrics. Model 3 achieved the strongest overall performance, attaining the highest accuracy (0.7678), specificity (0.8005), and ROC–AUC (0.8538). These results indicate that Model 3 provides the most robust discrimination between positive and negative cases, reflecting superior generalizability relative to the other configurations. Model 1 exhibited the highest recall (0.8303), demonstrating heightened sensitivity to positive cases, although this came with reduced specificity (0.6715), suggesting a greater propensity for false positives. Model 2 offered the most balanced behavior, achieving the highest F1 score (0.7851) along with competitive precision and specificity, reflecting an effective trade-off between sensitivity and precision.

Overall, the findings indicate that the architectural and regularization enhancements introduced in Model 3 substantially improve both discriminative ability and robustness. The progressive improvements across accuracy, specificity, and ROC–AUC highlight the benefit of stronger regularization in strengthening model stability and enhancing detection performance on the held-out test set.

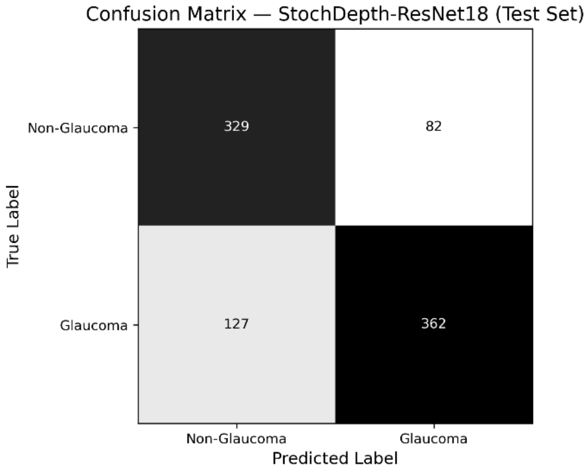

Fig. 2. Confusion Matrix of StochDepth-ResNet18 Model on Test Set

The confusion matrix in Fig. 2 summarizes the classification behavior of the StochDepth-ResNet18 model on the held-out test set. The model correctly identified 329 non-glaucoma cases and 362 glaucoma cases, demonstrating strong performance across both classes. The number of false negatives (127) is higher than the number of false positives (82), indicating that the model is slightly more conservative when predicting glaucoma; however, the true-positive count remains considerably larger than the false-negative count, reflecting effective sensitivity to glaucomatous RNFL patterns. Overall, the distribution of errors is consistent with the model's quantitative metrics and the confusion matrix confirms balanced detection capability across both healthy and glaucomatous eyes.

5.2 Ablation Study and Impact of Regularization Components

The ablation study across the three model configurations clarifies the individual and combined effects of the employed regularization strategies (Table 3). Model 1, incorporating only dropout, establishes a strong baseline and achieves the highest recall, indicating that dropout alone increases sensitivity by limiting overfitting and enabling the model to more readily identify positive cases. However, this comes at the expense of reduced specificity, suggesting a tendency toward higher false-positive rates.

Enhancing the architecture with stochastic depth in Model 2 improves both specificity and F1-score, demonstrating the stabilizing influence of randomly skipping residual blocks during training. This mechanism encourages the network to learn more resilient hierarchical features, leading to a more balanced trade-off between precision and recall.

Model 3, which adds Gaussian noise augmentation to the dropout–stochastic-depth configuration, achieves the highest accuracy, specificity, and ROC–AUC among all variants. These gains highlight the value of low-level noise injection in regularizing early convolutional responses, thereby increasing robustness to subtle variations in imaging conditions and reducing over-reliance on spurious correlations.

Taken together, the ablation results demonstrate that the progressive integration of dropout, stochastic depth, and Gaussian noise introduces complementary regularization effects. This combination substantially enhances feature robustness and improves the overall generalizability of the model for the target classification task.

Collectively, the progressive improvements across these ablations confirm that combining stage-level dropout, block-level stochastic depth, and pixel-level noise augmentation yields complementary regularization effects that substantially enhance the generalizability of ResNet-18 for medical imaging tasks such as glaucoma detection.

5.3 Training Dynamics of StochDepth-ResNet18

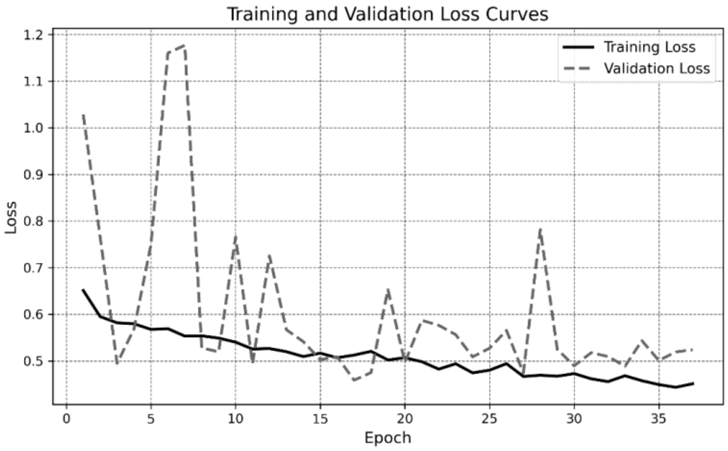

Fig. 3. Training v/s Validation Loss during training

Fig. 3 illustrates the training and validation loss trajectories over 50 epochs for the StochDepth-ResNet18 model. The training loss decreases smoothly and consistently, indicating stable optimization and effective learning of discriminative RNFL features. In contrast, the validation loss exhibits higher variance in the early epochs, reflecting sensitivity to stochastic regularization mechanisms (dropout, stochastic depth, and Gaussian noise) and the relatively small validation set size.

However, after approximately 15–20 epochs, the validation loss stabilizes around a narrower range and closely tracks the training loss. The absence of a sustained upward trend in validation loss suggests that the model does not overfit despite strong regularization, and the narrowing gap between training and validation curves indicates improved generalization as training progresses.

5.3.1 Fairness Evaluation across Demographic Groups

Table 4. Comparative Analysis of performance on Test Set of RNFLT maps with Race identities

Method	ES-Acc	Acc	ES-AUC	AUC	Asian AUC	Black AUC	White AUC	DPD	DE Odds
StochDepth ResNet18	**75.00**	76.78	**83.28**	85.38	85.39	81.06	85.11	13.33	8.31
FSCL+ w/ FIN [5]	74.04 ±0.67	77.02 ±0.81	78.69 ±0.55	**86.40** ±0.73	**88.39** ±1.13	81.17 ±0.43	**88.96** ±0.97	**10.67** ±0.88	**6.33** ±1.12
FIN [5]	73.03 ±1.45	**78.04** ±0.63	81.00 ±0.38	86.12 ±0.33	88.02 ±0.49	**82.15** ±0.26	86.26 ±0.81	17.26 ±1.95	15.22 ±0.74
Efficient Net + FIN [15]	NR	NR	81.4	84.5	85.2	81.6	84.3	NR	NR

Table 4 presents a comparative fairness evaluation of the proposed StochDepth-ResNet18 model against three state-of-the-art baselines: FSCL+ w/ FIN [5], FIN [5], and EfficientNet + FIN [15]. The proposed model achieves an ES-Acc of 75.00%, outperforming FSCL+ w/ FIN (74.04%) and FIN (73.03%), thereby indicating improved equity-scaled accuracy across demographic subgroups. In terms of overall accuracy, StochDepth-ResNet18 (76.78%) performs competitively with the FIN-based baselines, although FSCL+ w/ FIN (77.02%) and FIN (78.04%) report marginally higher values.

With respect to discrimination performance, the proposed model attains an ES-AUC of 83.28% and an overall AUC of 85.38%. These values compare favorably to EfficientNet + FIN [15] (ES-AUC = 81.4, AUC = 84.5) and remain close to the FIN-based benchmarks, where FSCL+ w/ FIN and FIN report overall AUCs of 86.40% and 86.12%, respectively. Across racial subgroups, StochDepth-ResNet18 demonstrates strong and well-balanced performance, achieving an Asian AUC of 85.39%, which is slightly lower than FSCL+ w/ FIN (88.39%) and FIN (88.02%) but higher than EfficientNet + FIN (85.2). For the Black subgroup, the proposed model obtains an AUC of 81.06%, comparable to Luo et al.'s reported values (81.17% for FSCL+ w/ FIN and 82.15% for FIN), and also similar to EfficientNet + FIN (81.6). Performance on the White subgroup (85.11%) is likewise competitive, closely matching EfficientNet + FIN (84.3) and falling within the range of values reported by FSCL+ w/ FIN (88.96%) and FIN (86.26%). Notably, EfficientNet + FIN does not report ES-Acc, overall accuracy, DPD, or DEOdds, and these omissions are explicitly marked as NR.

Regarding fairness-specific metrics, the StochDepth-ResNet18 model attains a DPD of 13.33% and a DEOdds of 8.31%. While these values indicate moderate group disparities, they remain within the variability range of existing FIN-based approaches. FSCL+ w/ FIN achieves the lowest disparities (DPD = 10.67%, DEOdds = 6.33%), whereas FIN exhibits substantially higher group differences (DPD = 17.26%, DEOdds = 15.22%).

Taken together, the results demonstrate that the proposed StochDepth-ResNet18 model offers fairness outcomes that are competitive with, and in several cases comparable to, established state-of-the-art methods. Importantly, it achieves this while maintaining strong and well-distributed diagnostic performance across Asian, Black, and White populations, underscoring its suitability for equitable medical AI applications.

6 Conclusion

This study introduced StochDepth-ResNet18, a regularization-enhanced architecture designed to improve the equity and reliability of glaucoma detection from OCT-derived RNFL thickness maps. By combining stochastic depth, residual-stage dropout, and Gaussian noise injection, the model is encouraged to learn stable structural representations rather than demographic-specific patterns. Across standard and equity-scaled evaluation metrics, the proposed approach demonstrated competitive diagnostic performance together with improved subgroup robustness, underscoring the value of architectural regularization for mitigating unintended bias.

Despite these promising findings, several limitations remain. All experiments were conducted on a single institutional dataset, necessitating further validation across diverse external cohorts to fully assess generalizability. Moreover, the proposed model does not incorporate explicit fairness-aware optimization and could benefit from integration with identity-adaptive modules—such as FIN—particularly in settings characterized by pronounced demographic imbalance.

Future work will explore multimodal data integration, domain generalization strategies, and prospective validation pathways to support safe and scalable clinical deployment. Overall, this work demonstrates that principled, minimally intrusive regularization techniques can provide a simple yet effective mechanism for enhancing both diagnostic accuracy and fairness in automated glaucoma screening systems.

References

1. Tham YC, Li X, Wong TY, Quigley HA, Aung T, Cheng CY. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology. 2014 Nov 1;121(11):2081-90. Author, F., Author, S.: Title of a

proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).

2. Hood DC, Kardon RH. A framework for comparing structural and functional measures of glaucomatous damage. Progress in retinal and eye research. 2007 Nov 1;26(6):688-710.
3. Muhammad H, Fuchs TJ, De Cuir N, De Moraes CG, Blumberg DM, Liebmann JM, Ritch R, Hood DC. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. Journal of glaucoma. 2017 Dec 1;26(12):1086-94.
4. Drukker K, Chen W, Gichoya J, Gruszauskas N, Kalpathy-Cramer J, Koyejo S, Myers K, Sá RC, Sahiner B, Whitney H, Zhang Z. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. Journal of Medical Imaging. 2023 Nov 1;10(6):061104-.
5. Luo Y, Tian Y, Shi M, Pasquale LR, Shen LQ, Zebardast N, Elze T, Wang M. Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. IEEE Transactions on Medical Imaging. 2024 Mar 18;43(7):2623-33.
6. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care?. AMA journal of ethics. 2019 Feb 1;21(2):167-79.
7. Noury E, Mannil SS, Chang RT, Ran AR, Cheung CY, Thapa SS, Rao HL, Dasari S, Riyazuddin M, Chang D, Nagaraj S. Deep learning for glaucoma detection and identification of novel diagnostic areas in diverse real-world datasets. Translational Vision Science & Technology. 2022 May 2;11(5):11-.
8. Maetschke S, Antony B, Ishikawa H, Wollstein G, Schuman J, Garnavi R. A feature agnostic approach for glaucoma detection in OCT volumes. PloS one. 2019 Jul 1;14(7):e0219126.
9. Luo Y, Shi M, Tian Y, Elze T, Wang M. Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning. InProceedings of the IEEE/CVF International Conference on Computer Vision 2023 (pp. 20471-20482).
10. NK J, Ali MH, Senthil S, Srinivas MB. Early detection of glaucoma: feature visualization with a deep convolutional network. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2024 Dec 31;12(1):2350508.
11. Pascal L, Perdomo OJ, Bost X, Huet B, Otálora S, Zuluaga MA. Multi-task deep learning for glaucoma detection from color fundus images. Scientific Reports. 2022 Jul 20;12(1):12361.
12. Bajwa MN, Malik MI, Siddiqui SA, Dengel A, Shafait F, Neumeier W, Ahmed S. Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning. BMC medical informatics and decision making. 2019 Jul 17;19(1):136.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
14. Huang G, Sun Y, Liu Z, Sedra D, Weinberger KQ. Deep networks with stochastic depth. InEuropean conference on computer vision 2016 Sep 17 (pp. 646-661). Cham: Springer International Publishing.
15. Shi M, Luo Y, Tian Y, Shen LQ, Zebardast N, Eslami M, Kazeminasab S, Boland MV, Friedman DS, Pasquale LR, Wang M. Equitable artificial intelligence for glaucoma screening with fair identity normalization. NPJ Digital Medicine. 2025 Jan 20;8(1):46.